Does Labelling Matter? The Effects of Labels on Responses – an Experience Sampling Study

Bachelor Thesis – Positive Psychology and Technology
Faculty of Behavioural Management and Social Sciences
University of Twente

Laura Suntrup

BSc. Psychology

Supervisor: Dr. Jannis T. Kraiss
Second supervisor: MSc., MRes. Mieke van Bergen
Date: 23.01.23

Abstract

Background: Previous studies examined the effect of labelling on responses regarding retrospective questionnaires. These studies have shown an association between response styles, such as the Extreme Response Style and the Net Acquiescence Response, and the effect of labelling on the responses. However, no study has investigated the effect of labelling in experience sampling method studies to date. Thus, this longitudinal study will further investigate the effect of fully-labelled compared to endpoint-only labelled Likert scales on response to individual negative affect item's distribution, mean leave, variability and inertia of affect.

Method: The study applied a mobile phone experience sampling method. Convenience sampling was used to recruit the participants. These (N = 47) were divided into two conditions: the fully-labelled condition (N = 24) and the endpoint-only labelled condition (N = 23). Despite the condition, each individual was supposed to complete ten semi-randomly scheduled ESM questionnaires per day for seven days in total. To test the distribution, the skewness and kurtosis values were calculated. Individual univariate multilevel models were performed for each negative affect item to check the mean level. To account for variability, four fixed effect models were conducted. The inertia was investigated through a linear mixed model.

Results: A significant interaction was found between the fully-labelled condition and *anxiety* $t(1774)=31.82$, $p<.001$ regarding inertia. The distribution and the mean level were comparable for the individual negative affect items. Also, the relationship between condition and variability was comparable for the items.

Conclusion: The findings conclude that different labelling in experience sampling method studies does not impact participants' responses. Nevertheless, the interaction between the condition and inertia was partially supported for item anxiety. Here, the response styles could have had an impact on the outcome. All in all, the study provided new insights into the effect of labelling on ESM responses.

Introduction

Just a beep from the phone and one gets reminded to fill in one's momentary mood. An hour later, the next beep strikes – This is a typical situation for a participant in an experience sampling method (ESM). ESM enables the researcher to capture individuals' self-reports to gain insides into their real-time feelings and experiences. This not only minimizes the retrospective memory bias and increases the ecological validity but has reformed how psychological research is carried out outside of the lab. Consequently, this led an increasing number of researchers to conduct ESM studies to gather within-person data (de Vries et al., 2001; Kraiss et al., 2022; Myin-Germeys et al., 2018; Scollon et al., 2003; Weermeijer et al., 2022). As a longitudinal design with its repeated measures within one sample, ESM guides in understanding the variability in experience of individuals in daily life and how the environment influences these through providing an examination of micro-level processes of an experience (Conner & Lehman, 2012; Csikszentmihalyi & Larson, 2014; Myin-Germeys et al., 2018). To assess the experiences of an individual, participants are asked to complete a questionnaire several times a day over a specific period of days concerning their mental state (Csikszentmihalyi & Larson, 2014; Myin-Germeys et al., 2018). ESM questionnaire items intend to assess the full emotional range, including the positive affect (PA) and the negative affect (NA). Further, it offers relevant outcomes, such as the dynamics of these affective states, which are seen as potential markers for psychopathology (Weermeijer et al., 2022). Inertia, for example, provides insights into how the negative affect state changed from one moment to the next, thus the affects resistance to change (Kuppens et al., 2010; Wichers et al., 2015). As a result of the aforementioned advantages, the use of ESM expanded in the past years; nonetheless, ESM still lacks research about its methodologies and design, which could, for example, allow greater replicability and further improve ESM (van Berkel et al., 2017; Weermeijer et al., 2022). Thus, it is relevant to investigate ESMs operationalization further.

When designing an ESM study, various parameters need to be considered. Whilst the study duration, sampling frequency and questionnaire length are defined as central design choices, questions such as how many response categories a rating scale should have and how the categories should be labelled are less considered by researchers (Myin-Germeys & Kuppens, 2022; Wetzel & Greiff, 2018). However, these design choices influence the participants' responses and, thus, the outcome and the psychometric properties (Myin-Germeys & Kuppens, 2022). Generally, response scales are different methodologies that support researchers in evaluating participants' judgments about, for example, their inner thoughts. They differ in the choice of anchor points or categories along a scale and the labelling of these anchor points (Lazovik & Gibson, 1984).

There are different response scale options for variables in ESM studies. For continuous variables specifically, continuous scales such as Visual Analogue Scales (VAS) and discrete scales such as Likert scales are prominent (Myin-Germeys & Kuppens, 2022). Likert scales are a simple measure to obtain broader values and attitudes (Johns, 2010). Considering that attitudes are seen as varying from negative to positive and can therefore be captured greatly by the Likert scale as it serves the purpose that individuals can express the strength and direction of their opinion about a topic (Garland, 1991; Johns, 2010). Compared to discrete scales, continuous scales have the advantage of choosing an answer between two response options. Nevertheless, both scale options are reliable and valid (Kuhlmann et al., 2017). Furthermore, ESM questions vary in the number of response categories. For example, when adjusting a Likert scale to one's research, one can include five response categories (5- point Likert scale).

Another important parameter for designing an ESM study is the labelling of the response categories. Mainly, labels can be seen as the description of the anchor point along the scale. Different labelling types exist, such as verbal, numerical, or only endpoint labels (Lazovik & Gibson, 1984; Wetzel & Greiff, 2018). To demonstrate, in the case of a 7- point Likert scale, one could choose to label all seven-response categories explicitly with numbers. Another option is to use verbal labels such as agreement statements of strongly disagree to strongly agree (full-labelling) or only to label the extremes such as the endpoint categories (endpoint-labelling) (Weijters et al., 2010). Further, the participant chooses a response by comparing the labelled categories (Weng, 2004). When only the extremes are labelled, for example, the researchers leave the other response categories open for interpretation.

Compared to characteristics like the number of response categories of Likert scales, which have been broadly explored in ESM studies, the labelling has been omitted (Weijters et al., 2013). This is remarkable as labels may influence all responses given and are influential for the response distribution (Weijters et al., 2010; Weijters et al., 2013). In non-momentary questionnaires, minor changes in the appearance of a rating scale are seen to influence the response outcome. Moreover, a fully-labelled scale is shown to be associated with more excellent reliability, and individuals seem to be more guided in interpreting the anchor points (Mateijka et al., 2016; Reips, 2010; Smyth et al., 2006; Weng, 2004). Menold et al. (2014), among others, underline the higher reliability of full-labelled scales compared to endpoint labelling. However, other studies state that endpoint labelling provokes greater reliability (Rodgers et al., 1992).

Furthermore, labels are seen to have an influence on the response styles of individuals. Respondents, for instance, either tend to select or avoid the extreme response categories (de Jong et al., 2008). Two response styles are essential for the stake of this research: the Net Acquiescence Response Style (NARS) and the Extreme Response Style (ERS).

The NARS can be seen as a bias that influences the respondents, despite the content of the items, to agree rather than disagree (van Herk et al., 2004; Weijters et al., 2010). According to a study by Weijters et al. (2010), who contrasted endpoint-only labelling with full labelling, the tendency to agree increases with fully-labelled scales. The study further suggested that this occurs due to the "clarity" of the full-labelled scale. For instance, in the case of ESM it could translate to an continues choice of either the positive or the negative side of the scale.

Conversely, the ERS captures the bias of respondents to choose the extreme response categories regardless of the content (Baumgartner & Steenkamp, 2001). Especially the spreading of the data is affected through ERS to extreme values (Baumgartner & Steenkamp, 2001; Greenleaf, 1992; Hurley, 1998). In contrast to the NARS, full labelling could lead to lower scores in ERS. The "anchor effect" is a pattern that could add to these findings. The research defines it as the individuals' tendency to choose the central response categories rather than the extreme response categories of a Likert scale (Bishop & Herron, 2015). Krosnick (1991, as cited in Weijters et al., 2010) and Swain et al. (2008, as cited in Weijters et al., 2010) explain this through the increased salience and attractiveness of the intermediate options. These studies underline that the effect of labelling and the influence of the response styles in retrospective questionnaires have been studied before. At the same time, it remains unknown what effect different types of labelling have on how people respond to ESM questions.

Emotional Variability and Inertia as two Examples of ESM Outcomes

Generally, emotional dynamics are relevant ESM outcomes to investigate. Humans' emotions continuously fluctuate as a response to internal and external events individuals face. There is a rich history of ESM research investigating further emotional fluctuation and regulation, namely emotional dynamics or affect dynamics (Dejonckheere et al., 2019; Houben et al., 2015; Larsen, 2000; Wichers et al., 2015). Eventually, it is also relevant to explore the effect of labelling on emotional dynamics outcomes as labelling may affect the respondents.

Especially the negative affect of the short-term emotional changes over time, emotional variability and inertia are two important outcomes that should be investigated further within the frame of this research. Emotional variability indicates the average emotional deviation from an individual mean positive or negative affect level over time (Dejockheere et al., 2019). Specifically, it refers to one's range of emotional experiences across time (Houben et al., 2015). Here, increased emotional fluctuation might be associated with decreased well-being and increased psychopathology (Houben et al., 2015; Thompson et al., 2017; van Zutphen et al., 2015).

Next, aforestated emotional inertia refers to the extent to which positive or negative affects carry across time (Dejockheere et al., 2019). The emotions are seen to be more self-predictive and opposing to change. This assumption suggests a widespread tendency to resist change and stay in a particular emotional state. The fact that the world is frequently observed and understood in ways consistent with the current emotional state contributes to this propensity (Dejockheere et al., 2019; Houben et al., 2015; Kuppens & Verduyn, 2017). Resultingly, the emotional experiences of an individual with a high level of emotional inertia are steadier over time.

Aim

In conclusion, it is relevant to consider the prior research regarding retrospective studies when investigating the effect of fully-labelling and endpoint labelling of Likert scales on the negative affect of the individual items in ESM studies. It is to assume that the outcomes are comparable when using similar design choices. Although Weijters et al. (2010) study regarding the NARS covers marketing research, it is predicted that also, in ESM questionnaires, respondents could have a higher tendency to agree within a fully-labelled condition. Therefore, they could have a higher tendency to choose continuously for example a negative affect item. This could be visible by a higher mean for participants in the fully-labelled condition. Additionally, through the ERS research, it is to predict that the response styles could also affect the emotional changes as a lower ERS in the fully- labelled condition leads to a lower standard deviation, resulting in a lower variability. Also, a higher tendency to agree (high NARS) could mean a lower standard deviation, leading to a lower variability. Furthermore, the "anchor effect" could influence the respondents to choose instead the central options of the fully-labelled scale, which leads to a high inertia as there is a steadier tendency of the responses. Finally, Dejonckheer & Erbas (2022) suggested that the design choices, such as labelling, influence the responses, which could be seen in a distribution difference, such as the kurtosis and the skewedness of the outcome. Specifically, ERS is seen to affect the descriptive statistical analyses as it is seen to increase the standard deviations and therefore affect the distribution towards the endpoints.

This research is investigating the two research questions: What is the effect of labelling the endpoints only compared to using a fully-labelled Likert scale on the response to individual negative affect items mean level and distribution? What is the effect of labelling the endpoints only compared to using a fully-labelled Likert scale on the response to individual negative affect item's variability and inertia of affect? Based on the aforementioned reasoning, following hypothesis are proposed:

H1: The distribution of the individual negative items in the endpoint only labelled condition will be to a greater extend skewed and have a lower kurtosis.

H2: The mean of the individual negative items will be higher in the fully-labelled condition

H3: The emotional variability will be lower in individual negative affect items of the fully-labelled condition

H4: The Inertia will be higher in the individual negative affect items of the fully-labelled condition

Methods

The Ethics Committee of Behavioral, Management and Social Sciences of the University of Twente approved the study (request number: 221244). For the research, the participants were divided into two conditions: the endpoint-only and fully-labelled conditions.

Participants

For this study, non-representative convenience sampling was utilised to gather participants. The group of respondents was contingent on their time and willingness to participate. Convenience sampling is a quick and cost-efficient sampling method which includes individuals that are available to the researcher and motivated to participate (Etikan et al., 2016; Stratton, 2021). Thus, the researchers involved personal contacts and a test subject pool (SONA) of the University of Twente.

Participants were required to possess English reading skills, be above the age of 18 and own a mobile phone, to be able to participate in this study. ESM studies, on average, include around 53 participants (van Berkel et al., 2017). This research was aimed at 60 participants, for each condition, 30 individuals. There were, in total, 40 participants in the endpoint-labelling condition and 48 participants in the fully-labelled condition. To avoid concerns regarding unrepresentative data due to a low response rate from participants, the arbitrary decision was made to exclude participants that responded to less than 33.3% of the questionnaires (Viechtbauer, 2022). After excluding participants who completed less than 1/3 of the ESM questionnaires, had missing data, or had too many data points, 47 remained (*N* = 47). Specifically, 23 participants were left in the endpoint-labelling condition with a mean age of 24.55 (*SD* = 8.29) and 24 participants in the fully-labelled condition with a mean age of 24.05 (*SD* = 7.31). Most participants were female, German, studying and finished High School (Table 1).

Randomisation

To prevent similarities in characteristics of participants due to convenience sampling, stratified randomisation was utilised. Therefore, a randomisation cluster was designed, separating the participants of the three-researcher collected from each other and the participants generated by SONA. Each researcher intended to recruit 20 participants, while SONA intended to collect around 60 participants. Next, the participants were grouped in blocks of six people each, where half of these individuals were divided into the first condition and the others into the second condition.

Design and Procedure

The research was conducted using the application Ethica, a commonly used tool for ESM studies (Ethica Data Services Inc, 2022). Each condition was pilot tested by the research team before the start of the study to avoid errors and to guarantee the correct set-up of the study. Before the start of the study, participants received a briefing email with relevant information about the study, their Ethica registration code (suitable to the condition) and information on the importance of filling out as many beeps as possible (see Appendix A). After the participants entered their Ethica registration code, they were given informed consent. Only when actively agreeing to the informed consent were participants able to start the study (see Appendix B).

The data collection started on the 7th of November 2022 at 7:30 and continued until the 13th of November 2022 at 22:30. The duration of 7 days is seen as “typical” for investigating the process within the individual (Dejonckheere & Erbas, 2022). Participants from the SONA system were credited 2 SONA credits when completing at least 1/3 of all questionnaires, and the other participants were not compensated. First, participants were asked to complete a baseline questionnaire that obtained their demographics and assessed their mental state. This questionnaire was available for the participants to complete throughout the entire study. Next to the baseline questionnaire, participants were asked to complete ten daily questionnaires for seven days, in total, 70 questionnaires. These so-called ESM questionnaires were triggered randomly at 90-minute intervals, and the participants were informed by a beeping sound from their phones. For seven days, the first 90- minutes interval started at around 7:30 and the last one at around 21:00. Each ESM questionnaire was available for 15 minutes before disappearing.

Measures

The research was part of a larger-scale study; therefore, only parts of the questionnaires included in the data collection were relevant to the research. Specifically, the demographics of the baseline questionnaire and the negative affect measures of the ESM questionnaire (see Appendix C & D).

Baseline Questionnaires

Demographics. The participants were asked for their age, gender, nationality, occupation, and education up to this point. Besides that, students at the University of Twente were asked if they signed up via SONA and their SONA reference number.

ESM questionnaires

Negative Affect Measure. The negative affect was measured with four items based on the Positive and Negative Affect Schedule (PANAS), a reliable measure for these affects (Watson et al., 1988). Common in ESM studies is a 7-point Likert scale offering the participant greater options for categorising their feelings (Eisel et al., 2022; Joshi et al., 2015). Therefore, a 7-point Likert scale was included for both conditions. The fully-labelled scale was continuously labelled with verbal labels and with numbers from “not at all” (1), to “very slightly” (2), “slightly” (3), “moderately” (4), “much” (5), “very much” (6), to “extremely” (7). The second condition, with the endpoint labels only, was numerically labelled and had verbal labels for the endpoints: “not at all” (1) and “extremely” (7). The four negative items were “*How anxious do you feel right now?*”, “*How irritable do you feel right now?*”, “*How down do you feel right now?*” and “*How guilty do you feel right now?*”. The fully-labelling condition for the negative affect showed a Cronbach’s alpha of α = .880. In the other condition, the negative affect a Cronbach’s alpha of α = .818.

Data Analysis

Pre-processing. In total, 36 participants were excluded due to the 1/3 (around 33.3%) cut-off point and an error in their Ethica app, which triggered too many questionnaires for the individuals. The fully-labelled condition was coded as 1, and the endpoint-only condition as 0. For hypothesis 3, the within-person standard deviation of each participant for the individual negative items was calculated in Microsoft Excel (Microsoft Corporation, 2018). For hypothesis 4, the lagged variable (t-1) within individuals was created in Microsoft Excel (Microsoft Corporation, 2018) for the beeps of each item. Here, the last response of each day was set as missing to prevent lags between values from separate days (Jans-Beken et al., 2018). Next, the lagged variables were included in an IBM SPSS version 28.0.1.0 (IBM Corp, 2022) dataset.

Statistical Analyses. To visualise the demographics of the respondents' descriptive statistics were analysed (Table 1). For the first hypothesis, the data was visualised to investigate the distribution of the individual negative affect items. To test for normal distribution, the Shapiro-Wilk test was performed. The confidence interval of the skewness and kurtosis values for each negative item and each condition were calculated. Here, the standard error of the descriptive was set as the basis for the confidence interval calculation. The second hypothesis required individual univariate multilevel models for each negative affect item. Viechtbauer (2022) underline that these mixed models are suitable for the interdependence of repeated observations per person. In each individual model, each negative item was set as the dependent variable, and the condition was set as the fixed effect.
The two conditions were compared by checking if the contrast was significant between the groups. For third hypothesis, four individual fixed effect model were run with the condition as fixed effect and the between-person standard deviation of each individual item as the dependent variable. For the last hypothesis, the autoregressive parameter (lagged variable: t-1) for each negative item was set as the predictor variable in a linear mixed model with the interaction between the group and the lagged variable. This parameter demonstrates the extent to which a previous state carries over into a current state (Jongerling et al., 2015). The variance-covariance structure AR1 (first-order autoregressive model at level 1) was specified to regress each observation upon the following observation (Jongerling et al., 2015). All calculations and analyses were made in IBM SPSS version 28.0.1.0 (IBM Corp, 2022) and Microsoft Excel (Microsoft Corporation, 2018). Further, for all analyses, the outcome was identified as significant in cases where the p-value was below .05.

Results

The endpoint-labelled condition consisted of 23 participants * 70 beeps = 1.610 measurements. Moreover, the fully-labelled condition consisted of 24 participants * 70 beeps = 1.680 measurements. The values of the outliers of the study were not extreme, therefore the outliers were included in the analysis. The average response rate of the ESM questionnaire and the baseline questionnaire was 58.76% (*SD* = 49.23%). The composite person means scores of the negative affect in the endpoint labelled condition (*M* = 1.92, *SD* = 1.09) was comparable to the fully-labelled condition (*M* = 1.88, *SD* = 0.91). Also, the composite scores of each individual item for each condition were comparable as well, for example the first items composite score for the fully-labelled condition was (*M* = 2.03, *SD* = 1.16), the endpoint only condition (*M* = 2.12, *SD* = 1.47).

Table 1

Summary of the Demographics: Age, Gender, Nationality, Occupation, and Education Level of the Participants Divided by the Conditions

Variable	Endpoint-only labelling	Full labelling
Age		
	24.55 (SD=8.29)	24.05 (SD=7.31)
	%	**%**
Gender		
Female	63.6	54.5
Male	31.8	40.9
Other	4.5	4.5
Nationality		
German	63.3	45.5
Dutch	27.3	45.5
Other	9.1	9.1
Occupation		
Working	18.2	18.2
Student	50.0	50.0
Studying/Working	31.8	22.7
Self-employed	-	4.5
Not working	-	4.5
Education		
Middle School	9.1	4.5
High School	63.6	45.5
Bachelor	18.2	40.9
Master	4.5	4.5
Other	4.5	4.5

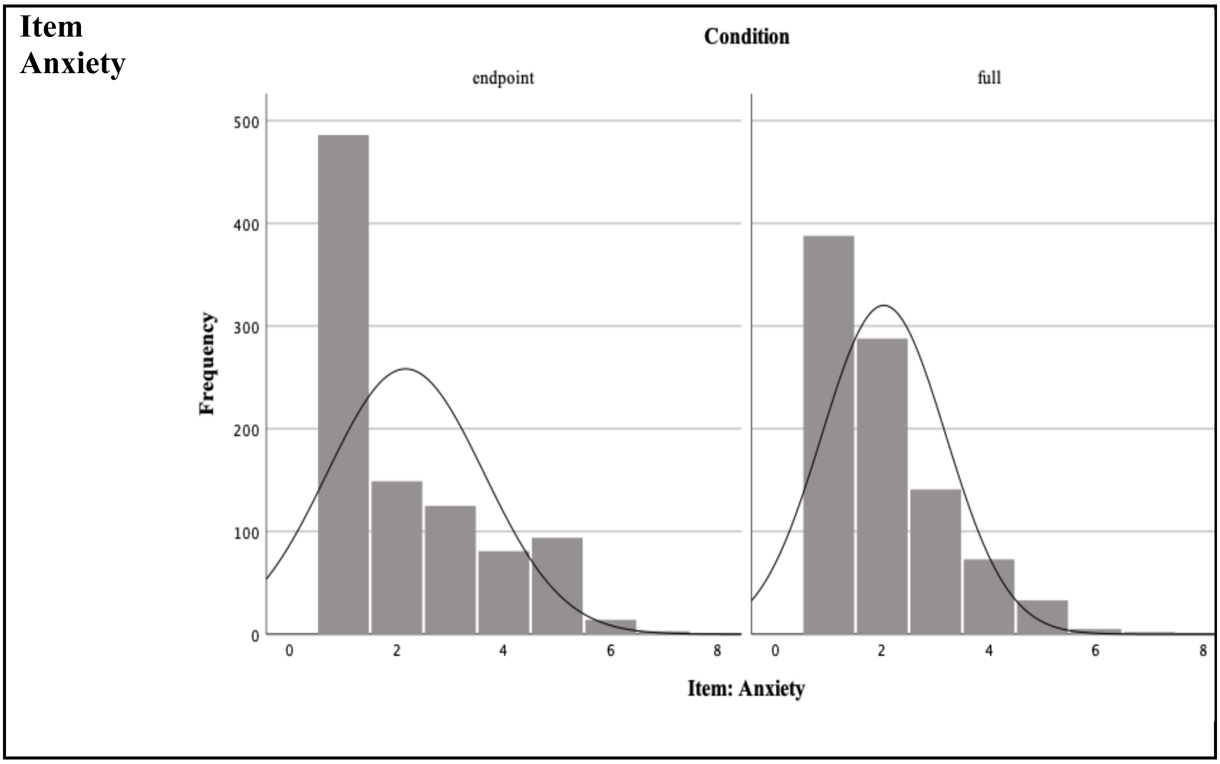
Visualization of the Distribution of the Negative Affect Items

First, a Shapiro-Wilk test was run to investigate if the data is normally distributed. When looking at the items individually it is noticeable that the data of all four items in both conditions significantly deviates from the normal distribution. The first items Shapiro-Wilk test outcomes *W*(947)=0.76, *p*<.001 for endpoint labelled condition and *W*(929)=0.81, *p*<.001 for the fully-labelled condition are comparable to the other items outcomes.

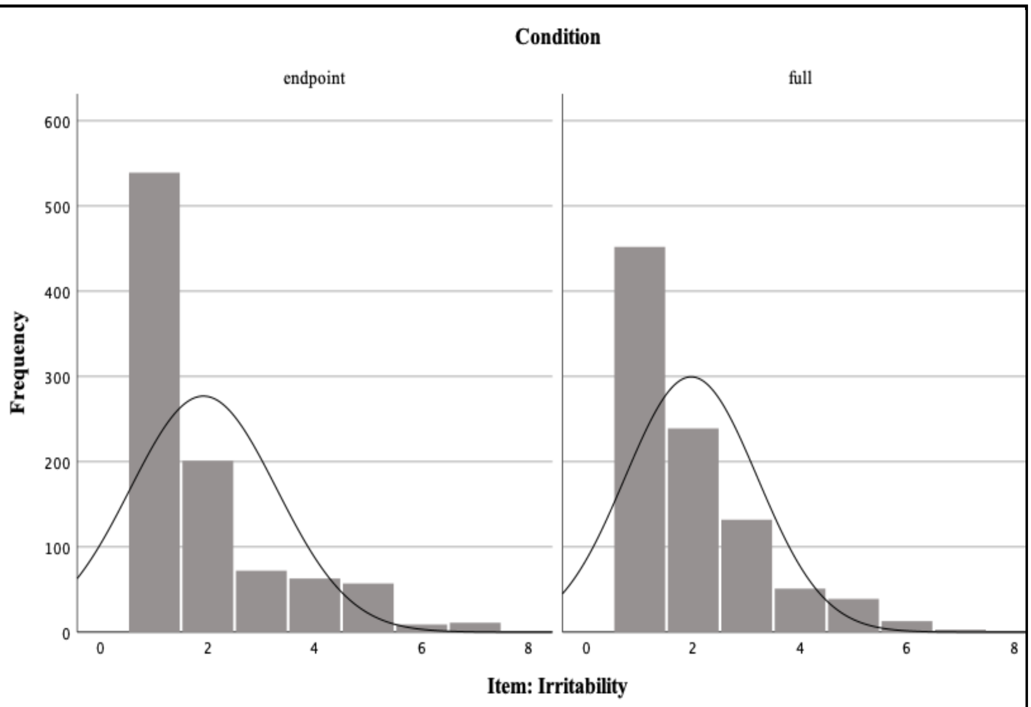
The distribution, specifically the kurtosis and skewness of all four items was comparable (Figure 1). All in all, there was an overlap in the confidence interval between both conditions in the kurtosis and skewness of all four items. For example, the skewness of the item *anxiety* was 1.03, 95% CI [-0.88, 1.19] in the endpoint labelling condition and in the fully-labelled condition 1.17, 95% CI [-1.01, 1.32]. This leads to the assumption that there is no significant difference between the conditions. Hence, the hypothesis that *endpoint labelled condition will be more skewed and have a lower kurtosis compared to the fully-labelled condition* in the individual items can be rejected.

Figure 1

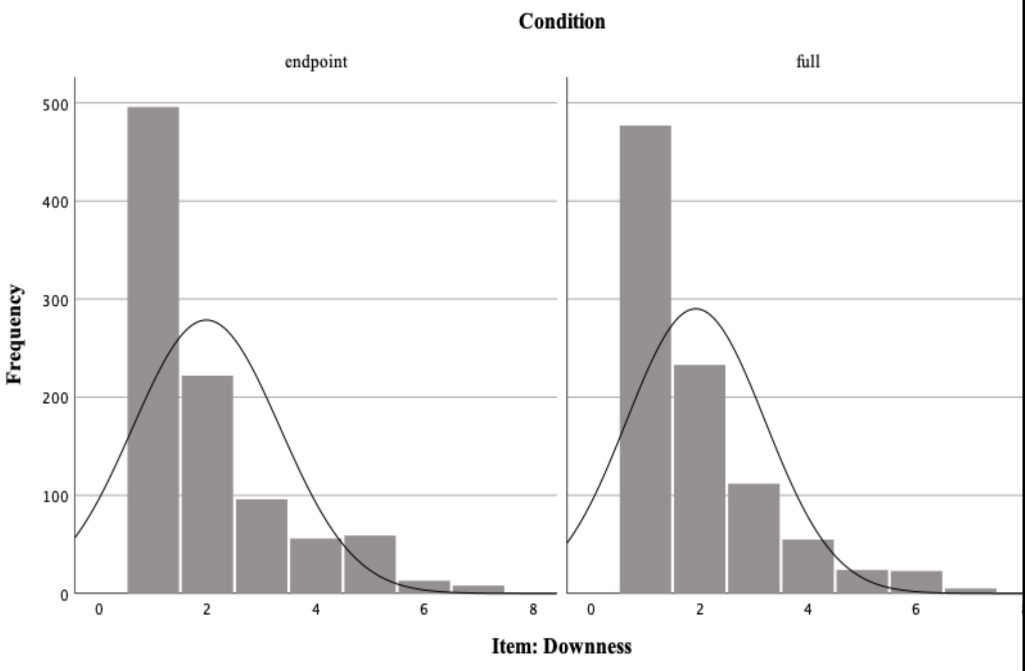
Distribution of the Negative Affect Item Anxiety, Irritability, Downness and Guiltiness Separated in Endpoint-only Condition and Fully-Labelled Condition



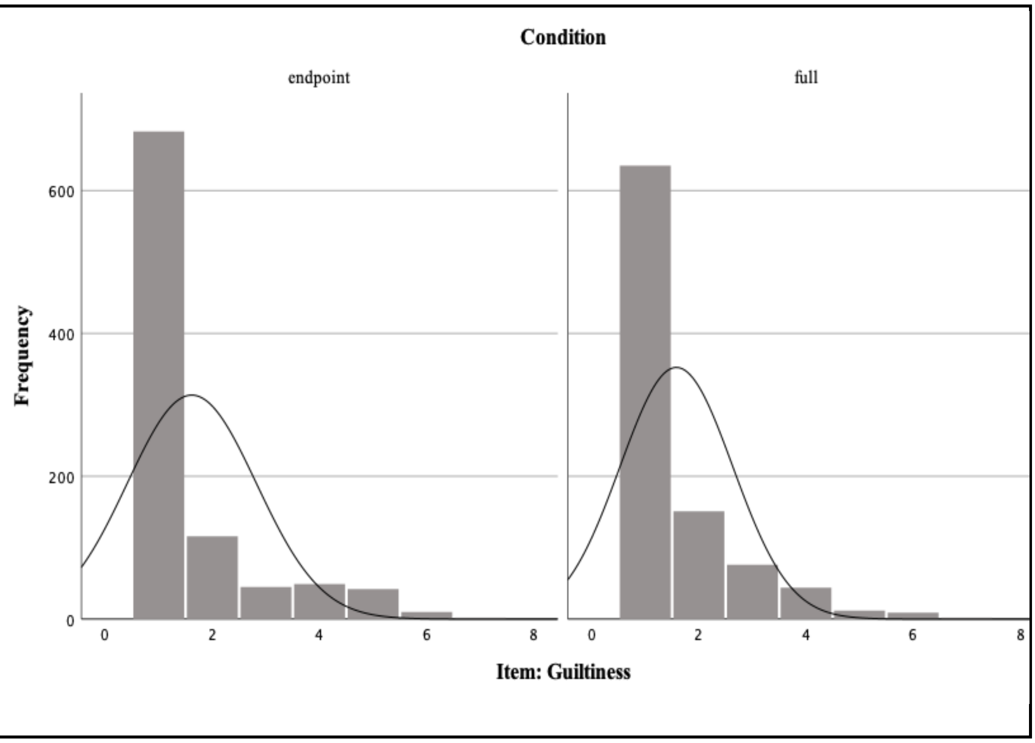
Item Irritability



Item Downness



Item Guiltiness



The Effect of Labelling on the Individual Items Mean Levels

Overall, the outcome of the analyses was comparable for the fully-labelled condition and the endpoint-only labelled condition for each individual negative item (Table 2). For instance, for *irritability* the estimated marginal mean of the endpoint labelled condition 1.92, 95% CI (1.21, 2.62) is comparable to the estimate for the fully-labelled condition 1.94, 95% CI (1.65, 2.23). Next to the comparable means, also the difference between the conditions for *irritability* was analysed as not significant *t*(45.705)= 13.3 *p*=.903. Resulting, the second hypothesis "*The mean of the individual negative items will be higher in the fully-labelled condition"* needs to be rejected. Here, the effect difference is not significant between the two conditions in each individual negative item.

Table 2

Scores of the Mixed-Effect Models between the Negative Affect Items and the Condition

Item	Parameter	*Estimate*	*SE*	*df*	*t*	Sig	Confidence Interval
Anxiety	Fully-labelled	2.01	0.18	49.829	11.33	<.001	[1.65, 2.36]
Anxiety	Endpoint only	2.13	0.43	49.979	11.79	.64	[1.26,2 .99]
Irritability	Fully-labelled	1.94	0.15	45.540	13.33	<.001	[1.65, 2.23]

Irritability	Endpoint labelled	1.92	0.35	45.705	13.21	.90	[1.21, 2.62]
Downness	Fully-labelled	1.94	0.16	45.418	12.13	<.001	[1.62, 2.27]
Downness	Endpoint labelled	2.01	0.39	45.581	12.41	.78	[1.23, 2.79]
Guiltiness	Fully-labelled	1.58	0.15	46.316	10.83	<.001	[1.29, 1.87]
Guiltiness	Endpoint labelled	1.57	0.35	46.469	10.8	.98	[0.86, 2.28]

Note. *df* Degrees of Freedom, *SE*= Standard Error

The Effect of Labelling on Affect Variability

The four negative affect items showed a comparable relationship between condition and variability (Table 3). The p-values in each model were not significant for the condition. Additionally, the R-Squared value for each negative item indicated that limited variance in the emotional variability can be predicted from the independent variable condition. For example item *anxiety* R2= .004, *F*(1, 45)=.185, *p*=.670 showed that 0.04% of the variance can be predicted by the condition fully-labelled or endpoint-only labelled. Concluding, the fully-labelled condition cannot show a lower variability as the condition did not affect the outcome. The third hypothesis "*The emotional variability will be lower in the fully-labelled condition*" can therefore be rejected.

Table 3

Scores of the Fixed Effect Model between Variability and the Conditions

Item	Variable	B	ß	*t*	p
Anxiety	(Constant)	.994		13.23	<.001
	Condition	-.045	-.064	-.43	.670
Irritability	(Constant)	1.049		12.10	<.001
	Condition	-.006	-.008	-.051	.959
Downness	(Constant)	.990		10.28	<.001
	Condition	.098	.108	.729	.470

Guiltiness	(Constant)	.692		.295	.770
	Condition	3.36	.151	1.023	.312

Note. Anxiety, R2 adjusted = .004 CI= Confidence Interval for B; Irritability, R2 adjusted = .000 CI= Confidence Interval for B; Downness, R2 adjusted = .000 CI= Confidence Interval for B; Guiltiness, R2 adjusted = .000 CI= Confidence Interval for B.

The Effect of Labelling on Inertia of Affect

Regarding the fourth hypothesis, it was investigated whether the condition influenced the effect. The results showed that the interaction between the lagged variable and the condition was a significant for the item *anxiety* $F(2, 1774)=5.88$, $p=.003$ and *downness* $F(2, 1771)=8.22$, $p<.001$. Interestingly, when looking at the two conditions for the item *anxiety* more closely it showed a significant interaction for the fully- labelled condition $t(1774)=31.82$, $p<.001$, but for the endpoint only labelled condition it resulted in an insignificant interaction $t(1774)=38.34$, $p=.754$. For *downness* both conditions showed a significant interaction effect, fully-labelled $t(1771)=29.96$, $p=.004$ and the endpoint-only labelled condition $t(1771)=29.89$, $p=.003$. By looking at the estimates one can determine in which direction this effect occurs. For the fully-labelled condition of the item *downness* 1.99, CI [1.82, 2.16] and for the endpoint labelled condition 1.99, CI [1.82, 2.16] there is no indicatable significant difference between the conditions. Leading to the conclusion, that the general interaction between the lagged variable and the condition influences downness, but the conditions itself show no significant difference. Furthermore, the items *irritability* $F(2, 1773)=.987$, $p=.373$ and *guiltiness* $F(2, 1768)=1.50$, $p=.223$, show an insignificant interaction between the lagged variable and the two conditions. Resulting, the fourth hypothesis "*The Inertia will be higher in the individual negative affect items of the fully-labelled condition"* can only partially be confirmed. Three out of the four items show no association between condition and effect and therefore the inertia is not higher in either of the conditions. *Anxiety* remains the only item that confirms the hypothesis as the fully-labelled condition resulted in higher inertia than the endpoint-only labelled condition.

Discussion

This study investigated the effect of labels on ESM questionnaires' individual responses. Previous retrospective questionnaire research has supported that different labelling affects the research outcomes (Mateijka et al., 2016; Reips, 2010; Smyth et al., 2006; Weng, 2004). However, no known study has investigated the effect of a fully-labelled scale compared to an endpoint-only labelled scale on the individual negative affect item's mean level and distribution, as well as the effect of these different labels on the individual negative affect item's variability and inertia of affect.

The Effect of Different Labels on the Individual Negative Items Means Level and Distribution

The results of the current study do not support previous research associating labelling with affecting the outcome and, therefore, the mean level and the distribution (Myin-Germeys & Kuppens, 2022; Weijters et al., 2013). Specifically, the results suggest no significant difference between the two conditions of endpoint-only labelling versus full labelling. This is contrary to the hypotheses that assumed, based on the reviewed literature, that labelling would affect the distribution and mean levels (Weijters et al., 2010). Specifically, the first hypothesis assumed that the distribution of the individual negative affects items in the endpoint-only condition would be, to a greater extent, skewed and have a lower kurtosis. At the same time, the second one expected the mean of the individual negative items to be higher in the fully-labelled condition. The hypotheses were mainly based on the two response styles, NARS and ERS. The NARS captures the tendency of individuals to agree rather than disagree with items despite their content. Weijters et al. (2010) pointed out that this primarily affects the fully-labelled scales. ERS, on the other hand, illustrates the bias to choose the extreme response options regardless of the content. Additionally, the "anchor effect" is relevant as it leads individuals rather to choose the central response options (Bishop & Herron, 2015).

Spratto et al.'s (2021) retrospective study suggested that endpoint-only labelling leads to higher ERS levels. However, this study substantiates that endpoint labelling is unlikely the only cause resulting in greater ERS and that there is still uncertainty about the strength of the effect. The results of this ESM research suggest that there is no effect of labelling by looking at the distribution, which was previously associated with being influenced by ERS. The retrospective study by Lau (2007) supports this finding and remarks that labelling response options might affect ERS levels to some extent but that this influence can also be situational or scale-specific. This could be a plausible explanation for why studies like Weijters et al. (2010) and Arce-Ferrer (2006) report an effect of labelling, specifically towards the endpoint conditions through ERS. However, this ESM study did not confirm these findings due to the possibility of having a different situational or scale-specific influence.

As previously established, ERS is seen to influence responses towards the endpoints. However, Weijters et al. (2013) findings suggest that the perceived intensity of the labels affects participants' response distribution. For example, if a respondent interprets an endpoint label as too "intense", they are less likely to choose it. This could explain the outcome of this study that labelling did not influence the individual negative affect items when being assumed on the basis of the response style ERS. The endpoint labels chosen in this research were "not at all" (1) and "extremely" (7), which could have been interpreted as too intense by the respondents. This could guide as an alternative explanation why the distribution was not affected by the endpoint-only labelled condition.

Regarding the NARS, the results of this ESM study contradict the research that the tendency to agree increases with the fully-labelled scales (Tourangeau et al., 2000; Weijters et al., 2010). Tourangeau et al. (2000) specifically supported that the "clarity" of the fully-labelled scale might be the reason for this. Therefore, the fully-labelled condition was assumed to show higher mean levels. The contradiction of the results could alternatively be explained by Spratto et al. (2021) who underlined that endpoint-only labelling may even present greater clarity about the meaning of the scale value to the individuals as verbal labels are seen as not as "precise" as numeric labels. They further explain that numeric values are less cognitively demanding and easier for individuals to memorize. Nevertheless, Krosnick and Fabrigar (2001, as cited in Spratto et al., 2021) suggest that this field still needs further research. Additionally, research states individuals' difficulty in matching their perceived feeling with a given response category (Hui & Triadis, 1989, as cited in Arce-Ferrer, 2006). This difficulty is perceived to a lesser extent when confronted with only endpoint labels, as there is greater room for interpretation. Yet, this study also did not support an effect of the endpoint condition on the outcome. All in all, for ESM this could mean that it does not matter which labelling is being used as it seemingly does not affect the responses.

The Effect of Different Labels on the Individual Negative Item's Variability and Inertia of Affect

Based on previous ESM research and assumptions based on the response styles, it was hypothesized that different labels influence the individual items' variability and inertia of affect (Bishop & Herron, 2015; Swain et al., 2008; van Herk et al., 2004; Weijters et al., 2010). Contrary to the hypothesis, the results indicated no significant difference between fully- labelled and endpoint-labelled scales for the variability. Here, it was suggested that the variability would be lower in the fully- labelled condition, which was not confirmed by the results—suggesting that the choice of labels for a scale in ESM does not affect the individual's emotional deviation from their mean level of negative affect across time. This is also contrary to Arce-Ferrer's (2006) study that underlined that ERS increases the standard deviations. Also, Tourangeau et al. (2007) mentioned that endpoint-only labelling would increase the use of response styles such as the ERS or the NARS as a form of heuristics. This study does not support these findings but suggests that in ESM studies, different labels do not increase the use of these response styles. In the definition of response styles, it is marked that both styles may affect the outcome despite the content of the items. It can be assumed that either the content of the items did affect the respondents, for example, to not choose the extreme response options or to a lower tendency to agree, or ESM questionnaires are differently affected by labelling than retrospective questionnaires.

Interestingly, cultural aspects also influence the response styles, which could be an alternative explanation. For example, individuals from Mediterranean and Latin American countries show greater ERS than individuals from other cultures (Arce-Ferrer, 2006). The average participant of this study is German, therefore, not residential in the prior stated counties. This may indicate that participants of this study show a slighter tendency to use the ERS.

The findings regarding the individual items' inertia of affect are different for the item *anxiety* compared to *irritability*, *downness* and *guiltiness*. This is suggesting that also the content of the items affected the participants. As previously defined, inertia refers to the extent to which, in this case, the negative effect carries over time. Based on the results, the inertia of the item's *anxiety* and *downness* were generally affected by the condition labelling. However, only anxiety acted as priorly assumed, suggesting that in the fully-labelled condition, the individual may carry over the negative affect *anxiety* to a greater extent than the endpoint-only condition. Generally, emotions fluctuate due to internal and external occurrences (Houben et al., 2015). Therefore, carrying over a negative affect from one moment to the next (high inertia) can be seen as a risk factor for, for example, future psychopathology (Wichers et al., 2015). However, the response styles could have influenced the individual's tendency to choose a specific response option, such as the central options ("anchor effect"), which could have indicated a steadier tendency of responses. Also, in the Spratto et al. (2021) research, the fully-labelled condition was assumed to have a higher tendency to select the midpoint response options due to the response styles, which resulted in being not supported by the studies outcomes. Based on this, the suggestion can be made that the response style might not have affected the results of this study and that more research needs to be conducted.

Strength and Limitations

The general study design was considered a strength of this research, as through longitudinal experience sampling design, it was possible to gather multiple measurements of participants' real-time feelings and experiences. Thereby, around 70 momentary self-reports of the individual's negative affect were collected within one week. Additionally, the ecological validity and the retrospective bias were reduced as ESM offers data collection in the individual's natural surroundings, and it questions their feelings when the beep is triggered on their mobile phone (van Berkel et al., 2018). Furthermore, the daily ESM questionnaires were not considered long, and thus, participants were not provoked to careless responses regarding the questionnaire length (Wetzel & Greiff, 2018).

However, four limitations should be deliberated in the interpretation of the findings. First, the operationalization of the ESM construct of the negative affect differs in various kinds of literature (Kirtley et al., 2019; Wetzel & Greiff, 2018). Thus, it is still being determined whether replication of the study with different operationalizations results in similar findings. Also, there is interindividual diversity in interpreting the labels of response categories (Wetzel & Greiff, 2018). Secondly, the study lacked a representative sample as mainly students (50% in both conditions, Table 1) participated in this study. Also, the mean age of the participants underlines this assumption. Thus, the sample is not representative of the general population and is limited in generalizability. Through the convenience sampling method, it can be assumed that the participating students are from the University of Twente and are, therefore, mostly familiar with the Likert scale design. This might have affected the results because students are generally more familiar with Likert scales and might automatically have filled in the blank response options of the endpoint-only labelled condition with their retrieved memory of a fully-labelled Likert scale (Arce-Ferrer, 2006; Wetzel & Greiff, 2018). Thirdly, the convenience sampling method could have had the effect that the researchers are familiar with some respondents. This could lead to the individuals giving socially desirable answers. Lastly, an unanticipated limitation is the low compliance of participants, 36 individuals had to be removed due to not completing sufficient daily questionnaires (<33.3%). Keeping all participants in could have negatively affected the data quality.

Implications and Future Research

The findings of this research suggest avenues and recommendations for future studies and implications. While previous research has focused chiefly on different labelling in retrospective questionnaires, no study has examined the effect of endpoint-only labelling compared to fully-labelled scales on the individual negative affect item's distribution and the variability and inertia of affect in an ESM study. Due to the findings of this study, new insights are contributed to the effect of labelling on ESM questionnaire responses. For instance, when investigating the item's inertia of affect, the item *anxiety* underlined a significant difference between the two conditions. This partially builds on existing evidence that the "anchor effect" could have influenced the individual to choose the central options, indicating a steadier response tendency (higher inertia).

The other results do not fit the theory that labelling affects the responses. Therefore, the implication for ESM studies is that using different labelling options does not seem to matter and therefore does not affect the participants' responses in these studies. Nevertheless, this statement is based on the two variations endpoint-only compared to fully labelled response options and the specific labels included in the study design, such as "Not at all" or "Extremely". However, as previously stated, the perceived intensity of labels is also relevant to consider (Weijters et al., 2013). For future research, it is thus to suggest conducting similar research with different variations of labels, such as "Very much" as the most positive option on a continuous scale compared to "Extremely" since it could be perceived as less extreme by participants. This also suggests experimenting further with different variations of Likert scales, which give possibilities for different verbal or numerical label combinations, for example, 6-point Likert Scales or 5-point Likert scales.

Moreover, there are specific suggestions for future research. Future research on the effect of labelling in ESM could include neutral labels compared to emotionally charged labels. Also, the level of specificity could vary from more general to specific labels. The population of this study is primarily students; therefore, it could be valuable to investigate the impact on different populations or cultures. Also, examining the effect of labelling on different types of experiences, like negative experiences, could be impactful. Furthermore, potential moderating effects could be considered, such as how participants' personalities, coping styles or individual differences, such as motivation, affect the relationship between labelling and responses. It could be of great interest to explore the potential implications further for real-world applications, for example, the clinical setting.

Additionally, ESM research could also further establish the effect of labelling on the response options using continuous scales instead of discrete scales. Matejka et al. (2016) investigated the impact of labelling on continuous scales; however, this effect was investigated using retrospective questionnaires. Therefore, one suggestion for future ESM research is to examine the effect of using continuous scales such as VAS and including different labels between points. This will allow participants to choose either a labelled option or an in-between labels option, which grants more room for expressing one's perceived feelings.

Lastly, for future studies, it is essential to consider how to engage participants in completing more questionnaires beforehand. Napa Scollon et al. (2009), for example, reported significantly improved compliance when participants were rewarded with money. This did not apply to the scope of this study, but future research could consider including rewards.

Conclusion

This study provided new insights into the effect of labelling on ESM responses. It disentangled the effect of fully-labelled compared to endpoint-only labelled Likert scales on the negative affect of *anxiety*, *irritability*, *downness* and *guiltiness* distributions, mean level, variability and inertia of affect. Previous study's findings regarding retrospective questionnaires were partially supported and discussed. The difference between the conditions regarding the item's distributions, mean level and variability was non-significant. The interaction between the condition and inertia was partially supported through a visible significant difference between the conditions of the item *anxiety*. This might be due to the response styles ERS, NARS and the "anchor effect". To further understand the effect of labelling, more longitudinal ESM studies needs to be conducted.

References

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style. Educational and Psychological Measurement, 66(3), 374−392.

Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156. https://doi.org/10.1509/jmkr.38.2.143.18840

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International journal of exercise science*, *8*(3), 297.

Chang, L. (1994). A Psychometric evaluation of four-point and six-point Likert-type scale in relation to reliability and validity. Applied Psychological Measurement, 18, 205−215.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology* (pp. 35-54). Springer, Dordrecht.

Davidson, R. J. (2003). Darwin and the neural bases of emotion and affective style. *Annals of the New York Academy of Sciences, 1000*(1), 316-336. https://doi.org/10.1196/annals.1280.014

De Jong, M. G., Steenkamp, J. B. E. M., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. Journal of Marketing Research, 45, 104−115.

de Vries, M., Caes, C., & Delespaul, P. (2001). The Experience Sampling Method in Stress and Anxiety Research. Anxiety Disorders, 289–306.

Dejonckheere, E., & Erbas, Y. (2022). Designing an Experience Sampling study.In I. Myin-Germeys, P. Kuppens, *The Open Handbook for Experience Sampling Methodology* (pp. 33-70). Retrieved from https://www.kuleuven.be/samenwerking/real/real-book/index.htm

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., Tuerlinckx, F. (2019). *Complex affect dynamics add limited information to the prediction of psychological well-being. Nature Human Behaviour, do*i:10.1038/s41562-019-0555-0

Eisel, G., Kasanova, Z., Houben, M. (2022). Questionnaire design and evaluation. In I. Myin-Germeys, P. Kuppens, *The Open Handbook for Experience Sampling Methodology* (pp. 71-90). Retrieved from https://www.kuleuven.be/samenwerking/real/real-book/index.htm

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136-151.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? Marketing Bulletin, 2, 66.

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research, 29*(2), 176-188. https://doi.org/10.1177/002224379202900203

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. Psychological Bulletin, 141(4), 901–930. https://doi.org/10.1037/a0038822

Hurley, J. R. (1998). Timidity as a response style to psychological questionnaires. *The Journal of Psychology, 132*(2), 201-210. https://doi.org/10.1080/00223989809599159

Jans-Beken, L., Jacobs, N., Janssens, M., Peeters, S., Reijnders, J., Lechner, L., & Lataster, J. (2019). Reciprocal relationships between state gratitude and high-and low-arousal positive affects in daily life: A time-lagged ecological assessment study. *The Journal of Positive Psychology*, *14*(4), 512-527.

Johns, R. (2010). Likert items and scales. *Survey question bank: Methods fact sheet*, *1*(1), 11-28.

Jongerling, J., Laurenceau, J. P., & Hamaker, E. L. (2015). A multilevel AR (1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate behavioral research*, *50*(3), 334-349.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British journal of applied science & technology*, *7*(4), 396.

Kirtley, O. (2018, December 4). The Experience Sampling Method (ESM) Item Repository. OSF. https://osf.io/kg376/

Kraiss, J. T., Kohlhoff, M., & ten Klooster, P. M. (2022). Disentangling between- and withinperson associations of psychological distress and mental well-being: An experience sampling study examining the dual continua model of mental health among university students. Current Psychology. https://doi.org/10.1007/s12144-022-02942-1

Kraiss, J. T., ten Klooster, P. M., Moskowitz, J. T., & Bohlmeijer, E. T. (2020). The relationship between emotion regulation and well-being in patients with mental disorders: A meta-analysis. Comprehensive Psychiatry, 102, 152189. https://doi.org/10.1016/j.comppsych.2020.152189

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology,* *5*(3), 213-236. https://doi.org/10.1002/acp.2350050305

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological science*, *21*(7), 984-991.

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22-26.

Larsen, R. J. (2000). Toward a science of mood regulation. *Psychological inquiry*, *11*(3), 129-141.

Lau, M. Y. K. (2008). *Extreme response style: An empirical investigation of the effects of scale response format and fatigue*. University of Notre Dame.

Lazovik, G. F., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. Applied Psychological Measurement, 8 (1), 49−57.

Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016, May). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5421-5432).

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales?. *Field methods*, *26*(1), 21-39.

Myin-Germeys, I., & Kuppens, P. (2022). The Open Handbook of Experience Sampling Methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies, 2nd edition. Independently published. Retrieved from https://www.kuleuven.be/samenwerking/real/real-book/index.htm

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry, 17(2), 123–132. https://doi.org/10.1002/wps.20513

Napa Scollon, C., Prieto, C. K., & Diener, E. (2009). Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being* (pp. 157-180). Springer, Dordrecht.

Reips, U. D. (2010). Design and formatting in Internet-based research. https://doi.org/10.1037/12076-003

Rodgers, W. L., Andrews, F. M., & Regula Herzog, A. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics-Stockholm-, 8*, 251-251.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Effects of using visual design principles to group response options in web surveys. *International Journal of Internet Science, 1*(1), 6-16.

Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2021). Seeing the forest and the trees: Comparison of two IRTree models to investigate the impact of full versus endpoint-only response option labeling. *Educational and Psychological Measurement*, *81*(1), 39-60.

Stratton, S. J. (2021). Population Research: Convenience Sampling Strategies. *Prehospital and Disaster Medicine*, *36*(4), 373–374. https://doi.org/10.1017/s1049023x21000649

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*(1), 116-131. https://doi.org/10.1509/jmkr.45.1.116

Thompson, R. J., Boden, M. T., & Gotlib, I. H. (2017). Emotional variability and clarity in depression and social anxiety. *Cognition and Emotion, 31*(1), 98-108. https://doi.org/10.1080/02699931.2015.1084908

van Berkel, N., Ferreira, D., & Kostakos, V. (2018). The Experience Sampling Method on Mobile Devices. ACM Computing Surveys, 50(6), 1–40. https://doi.org/10.1145/3123988

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346-360.

Van Zutphen, L., Siep, N., Jacob, G. A., Goebel, R., & Arntz, A. 2015). Emotional sensitivity, emotion regulation and impulsivity in borderline personality disorder: a critical review of fMRI studies. Neuroscience & Biobehavioural Reviews, 51, 64-76. https://dio.org/10.1016/j.neubiorev.2015.01.001

Viechtbauer, W. (2022). Statistical methods for ESM data. In I. Myin-Germeys, P. Kuppens, *The Open Handbook for Experience Sampling Methodology* (pp. 153-184). Retrieved from https://www.kuleuven.be/samenwerking/real/real-book/index.htm

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063. https://doi.org/10.1037/0022-3514.54.6.1063

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., ... & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*, 1-12.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236-247.

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The effect of familiarity with the response category labels on item response to Likert scales. *Journal of Consumer Research*, *40*(2), 368-381.

Weng, L. -J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. Educational and Psychological Measurement, 64(6), 956−972.

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales: Why we should think more carefully about the response format in questionnaires [Editorial]. *European Journal of Psychological Assessment, 34*(1), 1–5. https://doi.org/10.1027/1015-5759/a000469

Wichers, M., Wigman, J. T. W., & Myin-Germeys, I. (2015). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emotion Review*, *7*(4), 362-367.

Appendix A

ESM Briefing

1. General description of the experience sampling method (ESM)

What is ESM? You might not have heard of the term before. ESM is short for Experience sampling Method. This method is used to find out more about daily experiences such as how you feel, which activities you engage in and which people you meet during your day. Your role in this study will be to **fill out ten short questionnaires at different times throughout your day for one week**. The questionnaires only take about one minute to fill out, so they should not interrupt your daily activities too much. We want to measure life as it is, so it’s important that you live your normal life and not adjust your activities to this study. When you hear a beep, open your screen and fill out the short questionnaire immediately. If you only fill it out when you are on your own in a quiet environment, it won’t be informative as we are also interested in how you are doing when you are with others, when you are busy, when you are working etc. However, it’s okay if on some occasions you happen to miss a beep, we do not want you to adjust your life to the beep. That said, it is of course also important that you do not put yourself into a dangerous situation, such as filling out a questionnaire when you are driving or riding a bike. In general, try to fill it out as quickly as possible after the beep.

Please fill it out on as many occasions as you can.

Do you have any questions?

2. Install Ethica on the participant phone

Download the Ethica app via the play store or app store. Open the app and click on sign up if you don't have an account. If you do, then just log in.

App Store: https://appsto.re/i6h78DQ

Play Store: https://play.google.com/store/apps/details?id=com.ethica.logger

When you don't have an account first click on: You are a participant.

Here you are asked to fill in your first and last name, e-mail address that you would like to use and a password you can easily remember.

When you have done this, a screen will appear where you can enter your registration code for this study. You can find this in the e-mail we will send to you.

3. Additional information

a. Contact throughout the study

If you experience any issues with the app throughout the study or are stressed because of the study, do not hesitate to contact one of us, the researchers. Our email addresses and phone numbers are below this text. We will also send an email on day three of the study to check whether you have any questions.

Lorena Haase	l.haase@student.utwente.nl	+49 1573 2605744
Laura Suntrup	l.suntrup@student.utwente.nl	+49 1577 8980806
Gizem Elizabeth Konucu	g.e.konucu@student.utwente.nl	+31 6 53954282

b. Repeat the most important issues

Now, we will repeat the most important things you need to consider during the study

- Again, it is important that you keep up your normal daily routines. Do not cancel any appointments or make changes to your schedule due to the study
- Try to always carry your smartphone with you.

- Always answer the questionnaire immediately after the beep (of course without creating an unsafe situation).
- Please do not hesitate to get in touch with one of us if you have any issues with the app or experience distress due to the study.
- In case of discomfort and possible psychological distress, you can also contact the health services of the University of Twente. They can help you set up an appointment with one of the student psychologists via email info@campushuisarts.nl or phone +31 53 2030 204.

Appendix B

Informed Consent

Dear participant,

Thank you for your participation in this study.

Brief summary of project

The study is using the Experience Sampling Method (ESM) to obtain data. This means that 10 times a day there will be a prompt to answer a questionnaire containing about 20 items, which will take about 1 minute to complete. The questions regard your psychological well-being in the specific moment you are receiving the questionnaire and the time in-between questionnaires. It is important to fill out as many questionnaires as possible to ensure the success of the project.

To participate in this study, we need to ensure that you understand the nature of the research, as outlined in the participant information sheet. Please confirm at the bottom of the page to indicate that you understand and agree to the following conditions:

- I confirm that I have read the participant information sheet for this study. I have had the opportunity to consider the information, ask questions, and have had these answered satisfactorily
- I understand that to take part in this study, I should
  - Be at least 18 years old
  - Possess a basic level of English
- I understand that personal data about me will be collected for the purposes of the research study including age, gender, nationality, level of education, current studies, and primary occupation, and this data will be processed completely anonymous and in accordance with data protection regulations.
- I understand that taking part in this study involves that I will be filling in 10 questionnaires every day for one week.

- I am voluntarily taking part in this research, and I know that I can stop the research at any time without giving any reason, without my rights being affected
- I don't expect to receive any benefit or payment for my participation.
- I understand that I am free to contact the researchers or supervisor with any questions I may have in the future.
- I understand that the data collected in this study will be anonymized, and only be used for academic purposes i.e., writing a thesis for the bachelor and/or master.
- I understand that personal data that will be collected within this study will not be shared with anyone other than the study team.
- I agree to take part in this study.

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee/domain Humanities & Social Sciences of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-hss@utwente.nl

Study contact details for further information:

Name Researcher	***Email***
Laura Suntrup	l.suntrup@student.utwente.nl
Lorena Haase	l.haase@student.utwente.nl
Gizem Elizabeth Konucu	g.e.konucu@student.utwente.nl
Name Supervisor	***Email***
Jannis Kraiss	j.t.kraiss@utwente.nl
Thomas Vaessen	t.r.vaessen@utwente.nl

Appendix C

Baseline questions

Demographics

- Age: How old are you?
- Gender: What gender do you identify as? Male, female, other
- Nationality: What is your nationality? Dutch German Other
- Occupation: What is your current occupation? Student, Working, Self-employed, studying and working, not working, other

- Highest degree obtained: Middle school (such as MBO, MTS, MEAO or Haupt- oder Realschule), High school (such as HAVO, VWO, HBS or Gymnasium/ Berufsschule/ Berufskolleg), High school, Bachelor, Master, PhD, Other

Mental well-being (MHC-SF)

During the past month, how often did you feel...

1. Happy
2. Interested in life
3. Satisfied with life
4. That you had something important to contribute to society
5. That you belonged to a community
6. That our society is a good place or is becoming a better place, for all people
7. That people are basically good
8. That the way our society works makes sense to you
9. That you liked most parts of your personality
10. Good at managing the responsibilities of your daily life
11. That you had warm and trusting relationships with others
12. That you had experiences that challenged you to grow and become a better person
13. Confident to think or express your own ideas and opinions
14. That your life has a sense of direction or meaning to it
    a. Never
    b. Once or twice
    c. About once a week
    d. About 2 or 3 times a week
    e. Almost every day
    f. Every day

Anxiety (GAD-7)

Over the last two weeks, how often have you been bothered by the following problems?

1. Feeling nervous, anxious, or on edge
2. Not being able to stop or control worrying
3. Worrying too much about different things
4. Trouble relaxing
5. Being so restless that it is hard to sit still

6. Becoming easily annoyed or irritable
7. Feeling afraid, as if something awful might happen
    a. Not at all
    b. Several days
    c. More than half the days
    d. Nearly every day

Depression (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems?

1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could have noticed. Or the opposite being so fidgety or restless that you have been moving around a lot more

than usual

9. Thoughts that you would be better off dead, or of hurting yourself
    a. Not at all
    b. Several days
    c. More than half the days
    d. Nearly every day

Resilience (BRS)

Please respond to each item by marking one box per row

1. I tend to bounce back quickly after hard times
2. I have a hard time making it through stressful events.
3. It does not take me long to recover from a stressful event.
4. It is hard for me to snap back when something bad happens.
5. I usually come through difficult times with little trouble.
6. I tend to take a long time to get over setbacks in my life

a. Strongly disagree

b. Disagree

c. Neutral

d. Agree

e. Strongly agree

Perceived Stress (PSS)

The questions in this scale ask you about your feelings and thoughts during THE LAST MONTH. In each case, please indicate your response by placing an "X" over the circle representing HOW OFTEN you felt or thought a certain way.

1. In the last month, how often have you been upset because of something that happened unexpectedly?

2. In the last month, how often have you felt that you were unable to control the important things in your life?

3. In the last month, how often have you felt nervous and "stressed"?

4. In the last month, how often have you felt confident about your ability to handle your personal problems?

5. In the last month, how often have you felt that things were going your way?

6. In the last month, how often have you found that you could not cope with all the things that you had to do?

7. In the last month, how often have you been able to control irritations in your life?

8. In the last month, how often have you felt that you were on top of things?

9. In the last month, how often have you been angered because of things that were outside your control?

10. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

a. Never

b. Almost never

c. Sometimes

d. Fairly often

e. Very often

Cognitive reappraisal (ERQ subscale)

1. When I want to feel more positive emotion (such as joy or amusement), I change what I'm thinking about

2. When I want to feel less negative emotion (such as sadness or anger), I change what I'm thinking about.

3. When I'm faced with a stressful situation, I make myself think about it in a way that helps me stay calm

4. When I want to feel more positive emotion, I change the way I'm thinking about the situation

5. I control my emotions by changing the way I think about the situation I'm in

6. When I want to feel less negative emotion, I change the way I'm thinking about the situation.

1 Strongly disagree
2
3
4 Neutral
5
6
7 strongly agree

Rumination (CERQ subscale)

1. I often think about how I feel about what I have experienced.

2. I am preoccupied with what I think and feel about what I have experienced.

3. I want to understand why I feel the way I do about what I have experienced

4. I dwell upon the feelings the situation has evoked in me.

a. Almost never
b. Rarely
c. Occasionally
d. Frequently
e. Almost always

Acceptance (CERQ subscale)

1. I think that I have to accept that this has happened.
2. I think that I have to accept the situation.
3. I think that I cannot change anything about it.
4. I think I must learn to live with it.
    a. Almost never
    b. Rarely
    c. Occasionally
    d. Frequently
    e. Almost always

Appendix D
Daily questionnaire (ESM questionnaire)

Positive and negative affect

Below you can find several questions about your current feelings. Please try to indicate how you felt right before you started to answer the questionnaire!

- How *cheerful* do you feel right now?
- How *enthusiastic* do you feel right now?
- How *satisfied* do you feel right now?
- How *relaxed* do you feel right now?
- How *anxious* do you feel right now?
- How *irritable* do you feel right now?
- How *down* do you feel right now?
- How *sad* do you feel right now?
    - 1 (not at all) to 7 (extremely)

Perceived stress

- How stressed do you feel right now?
    - 1 (not at all) to 7 (extremely)

Stressful event + coping

Think of the most striking event or activity in last hour. How (un)pleasant was this event or activity?

- -3 (very unpleasant) to +3 (very pleasant)

How did you deal with this event?

- I kept thinking about it (rumination/savoring)
- I have tried to find a solution (active tackling)
- I tried to distract my attention from it (distraction)
- I expressed my emotions (emotion expression)
- I talked to others about it (social support seeking)
- I tried to look at it in a different way (positive/negative reappraisal)
    - Yes/no

Think of the most striking event or activity in the last hour. How stressful was this event or activity?

- 1 (not at all) to 7 (extremely)

Social context

Who are you with right now?

- Family member, friend, romantic partner, co-worker/fellow-student, unknown people/others, I am alone
- If not alone:
    - **I like this company**
    - 1 (not at all) to 7 (extremely)
    - **I would rather be alone**
    - 1 (not at all) to 7 (extremely)

Cognitive reappraisal

In the last hour, I tried to look at my problems from a different perspective

- 1 (not at all) to 7 (extremely)

Rumination

In the last hour, I have been thinking about my problems

- 1 (not at all) to 7 (extremely)

Acceptance

In the last hour, I could let go of my negative feelings without acting upon them

- 1 (not at all) to 7 (extremely)

Fully-labelled scale

1. Not at all
2. Very slightly
3. Slightly
4. Moderately
5. Much
6. Very much
7. Extremely